Gauge R&R Methods Compared

How do the ANOVA, AIAG, and EMP approaches differ?

Donald J. Wheeler

It would appear that there is still considerable confusion regarding which method to use in evaluating a measurement process. While there are many voices speaking on this subject, most of them fail to use the guidance provided by statistical theory, and as a result they end up in a train wreck of confusion and uncertainty. Here the three most common methods will be compared side by side.

Our example will use the data obtained from a gauge used to measure gasket thicknesses in mils. Three operators measure five parts two times each to obtain 30 measurements arranged into 15 subgroups of size 2 as shown in Figure 1. To understand these data it is helpful to observe that the operator-to-operator differences show up between the subgroups, the part-to-part differences show up between the subgroups, and the test-retest error shows up within the subgroups. This organization is the key to interpreting these data.

Operator	A					B					C				
Part	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
1st Value	167	210	187	189	156	155	206	182	184	143	152	206	180	180	146
2nd Value	162	213	183	196	147	157	199	179	178	142	155	203	181	182	154
Averages	164.5	211.5	185.0	192.5	151.5	156.0	202.5	180.5	181.0	142.5	153.5	204.5	180.5	181.0	150.0
Ranges	5	3	4	7	9	2	7	3	6	1	3	3	1	2	8
Variances	12.5	4.5	8	24.5	40.5	2	24.5	4.5	18	0.5	4.5	4.5	0.5	2	32

Figure 1: The Gasket Thickness Data

The comparison between the three approaches to gauge R&R studies will be carried out in two phases. In the first phase we shall look at the estimates of the various components of variation, and in the second we shall look at how the three approaches use these estimates to characterize the relative utility of the measurement system for a given application.

THE ANOVA APPROACH

The Analysis of Variance (ANOVA) approach to gauge R&R uses software to carry out the complex computations. This makes it a black-box technique. Once you have figured out how to enter the data above into the software correctly, the program will spit out an ANOVA table like that in Figure 2.

The large F-value for Parts (having a P-value less than 0.05) tells us that the measurement system can detect the differences between the five parts used in the study. This is good news because we want to be able to detect the part-to-part differences over and above measurement error.

Source	Sums of Squares	D.F.	Mean Squares	F-value	P-value
Parts	12791.1333	4	3197.78	262.11	0.0000
Operators	415.4000	2	207.70	17.02	0.0001
Interaction	103.2667	8	12.91	1.06	0.4392
Within	183.0000	15	12.20		
Total	13492.8000	29			

Figure 2: ANOVA Table for the Gasket Thickness Data

The large F-value for Operators tells us that there is a detectable operator effect present and the three operators are not measuring the parts the same. Specifically, the three operators do not all agree on the average value for the set of five parts. This is bad news. An operator effect is a nuisance component of measurement error. While the ANOVA approach detects this operator effect, it does not tell you which operator is out of line with the others.

The non-significant F-test for Interactions is another bit of good news. You do not want the operators to interact with the parts. If the operators measure the same parts differently you have a serious problem with your measurement process that will need to be fixed before the measurement system will be of any real utility.

ANOVA Repeatability: Our ANOVA estimate of the repeatability standard deviation or test-retest error has 15 degrees of freedom and is simply the square root of the Mean Square Within (MSW):

$$\text{Est. Repeatability} = \sqrt{\text{MSW}} = \sqrt{12.2} = 3.493 \text{ mils}$$

ANOVA Reproducibility: Since we have a detectable operator effect our ANOVA estimate of the Reproducibility standard deviation has 2 degrees of freedom and is found by using the Mean Square for Operators (MSO), the Mean Square for Interactions (MSI), and the Mean Square Within:

$$\text{Est. Reproducibility} = \sqrt{\frac{3}{30}[\text{MSO} - 2\,\text{MSI} - \text{MSW}]}$$

$$= \sqrt{0.1\,[207.7 - 2\,(12.91) - 12.2]} = 4.119 \text{ mils}$$

ANOVA Product Variation: If we think that the five parts used in this study are reasonably representative of the product stream we might use them to estimate the standard deviation of the product stream. This estimate would have 4 degrees of freedom and would be found using the Mean Square for Parts (MSP), the Mean Square for Interactions, and the Mean Square Within:

$$\text{Est. Product Standard Deviation} = \sqrt{\frac{5}{30}[\text{MSP} - 2\,\text{MSI} - \text{MSW}]}$$

$$= \sqrt{0.1667\,[3197.78 - 2\,(12.91) - 12.20]} = 22.948 \text{ mils}$$

ANOVA Total Variation: The three components of variation above may be combined to create an estimate of the standard deviation of the product measurements:

$$\text{Est. Std. Dev. of Product Measurements} = \sqrt{3.493^2 + 4.119^2 + 22.948^2} = 23.575 \text{ mils}$$

Since the computations for ANOVA can be extremely sensitive to round-off error, different software packages can give different results. I recently read a dissertation produced earlier this

year that had to compare the ANOVA results from different software packages because they gave different results. When black-box approaches disagree it is always hard to determine which result is correct if you cannot perform the computations required.

THE AIAG APPROACH

The Automotive Industry Action Group has been in the business of interpreting statistics for industry since 1990. Their approach to the data of Figure 1 consists primarily of computations of the estimates above and the computation of interpretative quantities. Their first step is to compute the upper range limit to check to see if any ranges from Figure 1 exceed this upper limit. Here the average range is 4.267, the upper range limit is 13.9, and all of the ranges are less than this value.

AIAG Repeatability: Step 2 is to compute an estimate of Repeatability using the average range. Here they divide by the traditional bias correction factor of $d_2 = 1.128$ to get an estimate with 13.4 degrees of freedom of:

$$Est.\ Repeatability\ =\ \frac{Average\ Range}{d_2}\ =\ \frac{4.267}{1.128}\ =\ 3.783\ \text{mils}$$

AIAG Reproducibility: Step 3 uses the operator averages to compute an operator range and then uses this to estimate the reproducibility standard deviation. The three operator averages are 181.0 mils, 172.5 mils, and 173.9 mils respectively. The range of these three values is $R_o\ =\ 8.5$ mils. The bias correction factor used by the AIAG manual here is the bias correction factor for estimating variances which is commonly known as d_2^*. For the range of three values this value is 1.906. The reproducibility estimate will have two degrees of freedom and is:

$$Est.\ Reproducibility\ =\ \sqrt{\frac{[Oper.\ Range]^2}{d_2^{*2}}\ -\ \frac{3}{30}\,(\text{Repeatability})^2}\ =$$

$$\sqrt{\frac{8.5^2}{1.906^2}\ -\ 0.10\,(3.783)^2}\ =\ 4.296\ \text{mils}$$

It should be noted that the AIAG approach computes the above number regardless of whether or not there is a detectable operator bias present in the data. While this value is appropriate in this case, this computation is inappropriate in the absence of a detectable operator bias.

AIAG Step 4 combines the Repeatability and Reproducibility:

$$Combined\ Repeatability\ \&\ Reproducibility\ =\ \sqrt{4.296^2\ +\ 3.783^2}\ =\ 5.724\ \text{mils}$$

AIAG Product Variation: Step 5 uses the averages for each part to compute a part-to-part range and then uses this range to estimate the standard deviation of the product stream. The five part averages are 158.0, 206.167, 182.0, 184.833 and 148.0 respectively. The range of these three values is $R_p\ =\ 58.167$ mils. The bias correction factor used by the AIAG manual here is the bias correction factor for estimating variances which is commonly known as d_2^*. This is done in spite of the fact that they are estimating a standard deviation. For the range of five values this value is 2.477. The product standard deviation estimate will have four degrees of freedom and is:

$$Est.\ Product\ Std.\ Dev.\ =\ \frac{Part\ Range}{d_2^*}\ =\ \frac{58.167}{2.477}\ =\ 23.483\ \text{mils}$$

AIAG Total Variation: Step 6 provides an estimate of the standard deviation of product measurements by combining the Repeatability, the Reproducibility, and the Product Variation:

$$\textit{Est. Std. Dev. of Product Measurements} = \sqrt{3.78^2 + 4.30^2 + 23.48^2} = 24.170 \text{ mils}$$

These estimates are compared with the estimates from other approaches in Figure 6 below.

THE EMP APPROACH

Unlike the two preceding approaches, the EMP approach begins with the average and range chart for the data of Figure 1. This allows you to quickly see the interesting aspects of your data before you get overwhelmed by computed values.
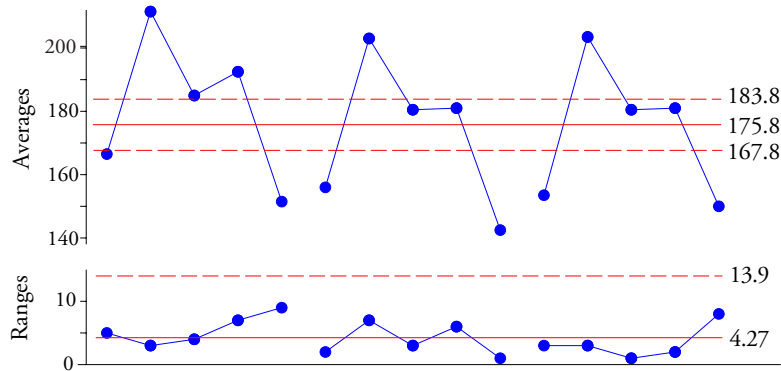

Figure 3: EMP Chart for Figure 1

In interpreting this chart it is important to understand that here we are using the average and range chart with experimental data. This means that we expect to see points outside the limits on the average chart—the whole point of doing an experiment is to create signals. At the same time, we do not want to see any signals on the range chart. Figure 4 shows what we look for on the EMP chart.
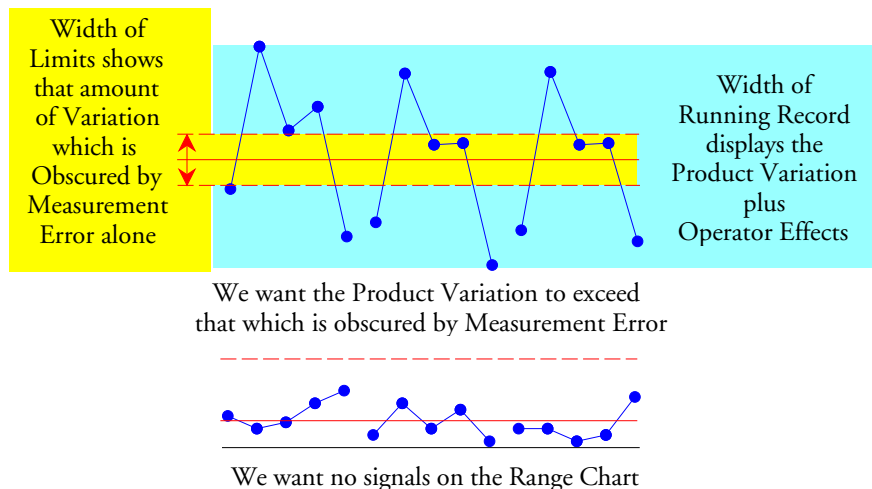

Figure 4: Interpreting the EMP Chart for Figure 1

This interpretation is due to the structure of these data. The operator-to-operator differences show up between subgroups. The part-to-part differences also show up between subgroups. The only source of variation left inside these subgroups is test-retest error, also known as repeatability. Thus, the limits on the average chart reflect that amount of variation which is obscured by measurement error. At the same time, the running record on the average chart shows the variation due to the product samples and the variation due to any operator differences that are present. Since we want to detect the product variation in spite of measurement error, we want to find points outside the limits on the average chart. The more points that are out, and the further out they are, the better we like it.

Since the range chart checks for consistency within the subgroups, any signals on the range chart would indicate that the measurement process was inconsistent. This check for consistency is missing from the ANOVA approach—it simply assumes that the measurement process is consistent. At least Step One of the AIAG method does check to see if there are any exceptional ranges. However, it is only the EMP approach actually constructs the range chart to show the consistency or inconsistency of the measurement process.

In addition to having points outside the limits on the average chart, and having the ranges all inside the limits, we also look for reasonable parallelism between the operators. We want them to agree on which parts are high and which parts are low. We also would like the operators to all have similar averages. In Figure 3 we have reasonable parallelism, but Operator A is a little bit high compared to the other two operators.

The reasonable parallelism in Figure 3 corresponds to the non-significant F value for the Interaction effect in the ANOVA table. The higher average for Operator A in Figure 3 is the operator effect found in the ANOVA table in Figure 2. How can we check to see if the operators all have the same average value for the five parts with EMP? We do this with an Analysis of Main Effects (ANOME) chart.

The Analysis of Main Effects is an ANOM technique explained in my February 2011 column "A Better Way to Do R&R Studies" The grand average is 175.8 and the average range is 4.267. We have $k = 15$ subgroups of size $n = 2$. The $m = 3$ operator averages are 181.0, 172.5, and 173.9. With an alpha-level of 5% the ANOME scaling factor is 0.592, giving limits of:

$$Grand\ Average \pm 0.592\ (Average\ Range)\ =\ 175.8 \pm 0.592\ (4.267)\ =\ 173.3\ to\ 178.3$$
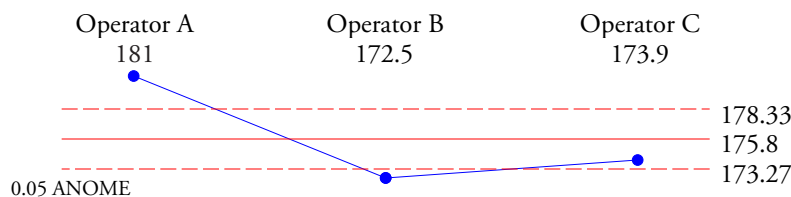

Figure 5: ANOME Chart for Operator Effect

And clearly Operator A is different from Operators B and C. Since this difference in operators represents a nuisance component of measurement error, we will want to estimate the effect this has upon the quality of the measurements. Thus, we will need to estimate the Reproducibility as well as the Repeatability in this case.

EMP Repeatability: In an EMP study several different choices for computational formulas are offered. The following are what I typically use. An unbiased estimate of the standard deviation of measurement error having 13.4 degrees of freedom is:

$$\text{Est. Repeatability} = \frac{\text{Average Range}}{d_2} = \frac{4.267}{1.128} = 3.783 \text{ mils}$$

EMP Reproducibility: For estimating the Reproducibility I use the same approach as Step 3 of the AIAG approach. Using the range of the three operator averages we end up with an estimate having 2 degrees of freedom:

$$\text{Est. Reproducibility} = \sqrt{\frac{[\text{Oper. Range}]^2}{d_2*^2} - \frac{3}{30}(\text{Repeatability})^2} =$$

$$\sqrt{\frac{8.5^2}{1.906^2} - 0.10\,(3.783)^2} = 4.296 \text{ mils}$$

EMP Product Variation: For estimating the product standard deviation we use a formula that is slightly more precise than the one used in the AIAG study. The range of the five part averages is still 58.167, and the bias correction factor is still 2.477, but now we remove some bias that was ignored earlier. So, using the range of the part averages we obtain an estimate having 4 degrees of freedom:

$$\text{Est. Product Std. Dev.} = \sqrt{\frac{[\text{Part Range}]^2}{d_2*^2} - \frac{5}{30}(\text{Repeatability})^2} =$$

$$\sqrt{\frac{58.167^2}{2.477^2} - 0.1667\,(3.783)^2} = 23.433 \text{ mils}$$

EMP Total Variation: As before, we combine the three estimates above to obtain an estimate of the standard deviation of the product measurements is:

$$\text{Est. Std. Dev. of Product Measurements} = \sqrt{3.78^2 + 4.30^2 + 23.43^2} = 24.121 \text{ mils}$$

The table in Figure 6 compares the estimates of these various quantities obtained from the three approaches. There is no practical difference between the various estimates of these four fundamental components of variation.

Component	ANOVA	AIAG	EMP
Repeatability (Est. Std. Dev. of Measurement Error)	3.49	3.78	3.78
Reproducibility (Est. Std. Dev. of Operator Effect)	4.12	4.30	4.30
Product (Est. Std. Dev. of Product Stream)	22.95	23.48	23.43
Product Measurements (Est. Std. Dev. of Observations)	23.57	24.17	24.12

Figure 6: Comparison of Estimates Across Approaches

The difference between the three approaches does not lie in the values found in Figure 6, but rather lies in how these estimates are used to describe the relative utility of the measurement system.

VARIANCE COMPONENTS FOR ANOVA

The quantities in Figure 6 are estimates of the square roots of the variance components. In the ANOVA approach we square the estimates from the ANOVA column of Figure 6 and express the first three numbers as a percentage of the last of these squared values to obtain a table of variance components as shown in Figure 7. Since the last component is the sum of the first three components, it is customary to express each of the first three components as a percentage of the last component.

Variance Component	ANOVA Estimates	Percentages
Repeatability (Measurement Error)	12.200	2.19%
Reproducibility (Operator Effect)	16.968	3.05%
Product (Product Stream)	526.627	94.75%
Product Measurements (Observations)	555.795	

Figure 7: Variance Components for ANOVA Approach

ANOVA Relative Utility: Once we have the estimates found above we will want to use them to obtain a statistic that will characterize the relative utility of the measurement system for measuring this product. The standard statistic for this purpose is the estimate of the *Intraclass Correlation Coefficient*. This statistic is found in Figure 7:

$$\textit{Estimated Intraclass Correlation} \;\; = \;\; \frac{526.6}{555.8} \;\; = \;\; 0.9475$$

This value is properly interpreted to mean that the variation in the product stream accounts for 94.75 percent of the variation in the product measurements. Conversely, the repeatability and reproducibility combine to account for 5.25 percent of the variation in the product measurements.

The Intraclass Correlation Coefficient defines that percentage of the variation in the product measurements that is directly attributable to variation in the product stream. This fact makes it easy to explain in practice. The fact that it has a fine ancestry of high-brow mathematical statistics dating back to the 1920s makes it the traditional measure of relative utility of a measurement system for a given application. Accept no substitutes.

VARIANCE COMPONENTS FOR EMP APPROACH

With the EMP approach we end up with the essentially the same estimates of the standard deviations for Repeatability, Reproducibility, and Product Variation as obtained from the other two approaches. As with the ANOVA approach we can square the values in the EMP column of Figure 6 to create the table of variance components for the EMP approach seen in Figure 8. Once again, we express each of the first three values as a percentage of the last value.

Here our estimate of the Intraclass Correlation Coefficient is 94.37 percent. So we estimate that about 94 percent of the variation in the product measurements is attributable to the variation in the product stream, and less than 6 percent is due to repeatability and reproducibility. Thus, the EMP approach yields an estimate of the Intraclass Correlation Coefficient that is essentially identical to the value found using the ANOVA approach. Even though the numbers are slightly

different, the overall interpretation remains the same with EMP as with the ANOVA approach.

Variance Component	EMP Estimates	Percentages
Repeatability (Measurement Error)	14.311	2.46%
Reproducibility (Operator Effect)	18.456	3.17%
Product (Product Stream)	549.105	94.37%
Product Measurements (Observations)	581.872	

Figure 8: Variance Components for EMP Approach

Thus, for measuring this product, this measurement system is already very good, and the impact of the operator difference is minimal. (If we are able to get Operator A to operate in line with Operators B and C, then the reproducibility component would go to zero, and the Intraclass Correlation Coefficient would increase from 94.4 percent to around 97.5 percent.)

THE AIAG "PROPORTIONS"

The AIAG approach does not concern itself with the components of variance. Instead, It computes its own measures of relative utility by using the quantities from Figure 6 directly. The AIAG approach computes ratios by dividing each of the first three AIAG values in Figure 6 by the last AIAG value, 24.17. These values are then multiplied by 100 to obtain the AIAG "percentages" shown in Figure 9. Then the repeatability and reproducibility percentages are compared to guidelines to determine the relative usefulness of the measurement system.

The AIAG manual interprets the values in Figure 9 to mean that Repeatability consumes 16% of the total variation, Reproducibility consumes 18% of the total variation, and the product stream consumes 97 percent of the total variation. Of course, the immediate problem with the three "percentages" in Figure 9 is that they add up to 130%!

Source	AIAG Estimates	Percentages
Repeatability (Measurement Error)	3.783	15.65%
Reproducibility (Operator Effect)	4.296	17.77%
Product (Product Stream)	23.483	97.15%
Product Measurements (Observations)	24.171	

Figure 9: The AIAG "Proportions"

Not knowing what else to do about this problem, the AIAG manual simply inserts a statement at this point to the effect that "The sum of the percent consumed by each factor will not equal 100%." This statement has no explanation attached. There is no guidance offered on how to proceed now that common sense and every rule in arithmetic have been violated. Just a statement that these numbers do not mean what they were just interpreted to mean, and the user is left to his or her own devices. Unfortunately, unlike the Red Queen (or the AIAG), when it comes to arithmetic we do not get to say that things mean whatever we want them to mean.

To make sense of the AIAG "percentages" we have to construct a couple of right triangles as shown in Figure 10. There the various estimates obtained by the AIAG approach are shown as the sides of the triangles.
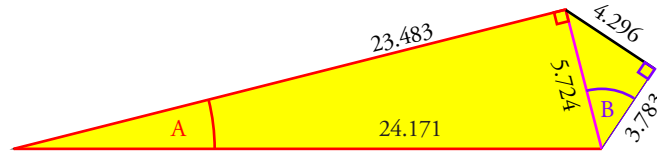
Figure 10: The Quantities Used by AIAG to Compute Pseudo-Percentages

Those who are familiar with high school trigonometry will immediately recognize that the ratio of Repeatability (3.783) to Total Variation (24.171) is the sine of angle A times the cosine of angle B.

$$AIAG\ Repeatability\ =\ \frac{3.783}{24.171}\ =\ \frac{5.274}{24.171}\ \frac{3.783}{5.274}\ =\ (\ Sine\ A\ )\ (\ Cosine\ B\ )\ =\ 0.1565$$

Likewise the ratio of Reproducibility (4.296) to Total Variation is seen to be the sine of angle A times the sine of angle B.

$$AIAG\ Reproducibility\ =\ \frac{4.926}{24.171}\ =\ \frac{5.274}{24.171}\ \frac{4.926}{5.274}\ =\ (\ Sine\ A)\ (\ Sine\ B\ )\ =\ 0.1777$$

The ratio of Product Variation (23.483) to Total Variation is the cosine of angle A.

$$AIAG\ Product\ Variation\ =\ \frac{23.483}{24.171}\ =\ (\ Cosine\ A)\ =\ 0.9715$$

So, in interpreting these ratios as percentages the AIAG group is effectively ignoring the Pythagorean theorem and confusing trigonometric functions with proportions. The fact that these numbers are irredeemable nonsense may be seen by comparing them with the true proportions given in Figure 11. The Repeatability consumes about 2.2 percent of the total variation rather than the 16 percent erroneously suggested by Figure 9. The Reproducibility consumes about 3.0 percent of the total variation, rather than the 18 percent erroneously suggested by Figure 9.

Component	ANOVA	AIAG	EMP
Repeatability (Measurement Error)	2.2%	15.7%	2.4%
Reproducibility (Operator Effect)	3.0%	17.8%	3.2%
Product (Product Stream)	94.8%	97.1%	94.4%
Total Variation Explained	100.0%	130.6%	100.0%

Figure 11: Measures of Relative Utility from the Three Approaches

So what can you learn from the AIAG gauge R&R study? Virtually nothing that is true, correct, or useful. You have taken the time and gone to the trouble to collect good data, and have even obtained reasonable estimates for the various components of variation, then you have wasted the information obtained by performing hopelessly flawed computations that have no meaningful interpretation. For more on this subject see my column of January 2011, "Problems with Gauge R&R Studies."

SUMMARY

So while all three approaches start off with essentially the same estimates, only the ANOVA approach and the EMP approach give theoretically sound, easy to interpret, and useful results.

The AIAG approach simply overstates the damage due to measurement error and condemns the measurement process. While both the ANOVA approach and the EMP approach show this measurement system to have very good utility for measuring this product, the AIAG approach erroneously suggests that the combined R&R consumes 24 percent of the total variation, and that as a consequence, this measurement system is of "marginal" utility. So who should you trust when evaluating your measurement processes? Those who flunked high school trigonometry, or those who build their approaches on sound statistical theory?